JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 9, Number 1, 2002
Mary Ann Liebert, Inc.
Pp. 35–54

Protein Engineering Study of Protein L by Simulation

JON M. SORENSON[1,3] and TERESA HEAD-GORDON[2,3]

ABSTRACT

We examine the ability of our recently introduced minimalist protein model to reproduce experimentally measured thermodynamic and kinetic changes upon sequence mutation in the well-studied immunoglobulin-binding protein L. We have examined five different sequence mutations of protein L that are meant to mimic the same mutation type studied experimentally: two different mutations which disrupt the natural preference in the β-hairpin #1 and β-hairpin #2 turn regions, two different helix mutants where a surface polar residue in the α-helix has been mutated to a hydrophobic residue, and a final mutant to further probe the role of nonnative hydrophobic interactions in the folding process. These simulated mutations are analyzed in terms of various kinetic and thermodynamic changes with respect to wild type, but in addition we evaluate the structure–activity relationship of our model protein based on the ϕ-value calculated from both the kinetic and thermodynamic perspectives. We find that the simulated thermodynamic ϕ-values reproduce the experimental trends in the mutations studied and allow us to circumvent the difficult interpretation of the complicated kinetics of our model. Furthermore, the level of resolution of the model allows us to directly predict what experiments seek in regard to protein engineering studies of protein folding—namely the residues or portions of the polypeptide chain that contribute to the crucial step in the folding of the wild-type protein.

Key words: phi-values, off-lattice models, protein L, protein folding, multiple histogram method.

INTRODUCTION

While the experimental effort in structural genomics is partly focused on providing new fold classifications, computation and theory should play a complementary role of contributing structural, kinetic, and thermodynamic information across whole genomes. However, in order to pursue a computationally feasible protein model of genomic-scale scope that has *predictive value* a series of validation steps are required. We have now completed multiple studies at a simplified level of description of protein folding addressing issues of protein sequence design (Sorenson and Head-Gordon, 1999), the role of solvation and interaction complexity in protein folding models (Sorenson and Head-Gordon, 1998), the ability to

[1]Department of Chemistry, University of California, Berkeley, Berkeley, CA 94720.
[2]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720.
[3]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

design and validate the folding of complex topologies (Sorenson and Head-Gordon, 2000a), as well as longer protein chains (Sorenson and Head-Gordon, 2002). We believe that they, in combination, indicate the utility of this level of modeling as a productive step forward given current computational limits on the feasibility of genomic-scale modeling.

In this next step of validation, we examine the ability of our minimalist model to reproduce experimentally measured thermodynamic and kinetic changes upon sequence mutation in the well-studied IgG-binding protein L (Gu *et al.*, 1997; Kim *et al.*, 1998; Kim *et al.*, 2000; Gu *et al.*, 1999). This is a particularly important validation step as we propose to predict the consequence of sequence mutations, as either benign to folding rates and stability, or as involving residues that are critical for native state formation, that begins to cover the diversity in sequence space that is part of the challenge of genomic-scale structural biology and proteomics. We have examined five different sequence mutations that are meant to mimic the same mutation type studied experimentally: two different Gly $\rightarrow$ Ala mutations of protein L studied by Gu *et al.* (1997, 2000) which disrupt the natural preference in the β-hairpin #1 and β-hairpin #2 turn regions for conformations not sterically accessible to nonglycine residues, and two different helix mutants that mimic the Glu $\rightarrow$ Ile (E32I) helix mutant examined by Kim *et al.* (1998, 2000) and Gu *et al.* (1999), where a surface polar residue in the α-helix has been mutated to a hydrophobic residue. We also examined a fifth mutant constructed by changing a neutral bead to a hydrophobic one in the first β-hairpin turn to further probe the role of nonnative hydrophobic interactions in the folding process, another possible consequence of the experimental Gly $\rightarrow$ Ala mutation in the turn region as described by Gu *et al.* (1997) and Kim *et al.* (2000).

These simulated mutations are analyzed in terms of various kinetic and thermodynamic changes with respect to wild type, but in addition we evaluate the structure–activity relationship of our model protein based on the ϕ-value (Matouschek *et al.*, 1984).

$$\phi = \frac{-RT \ln(k_{mut}/k_{wt})}{\Delta\Delta G^0}. \tag{1}$$

Protein folding experiments measure the mutant and wild-type folding rates, k_{mut} and k_{wt}, as well as $\Delta\Delta G^0$, the change in native-state stability after the mutation is made, to obtain indirect knowledge of the role of that residue in forming the transition-state.

We have considered two different ways to evaluate ϕ-values in order to match and validate our model against experimental mutations on protein L. For the five mutations examined here, we evaluated ϕ-values based on the kinetic definition, but our results indicate that the kinetic analysis based on Equation (1) is inappropriate because our model landscape is overly frustrated (Nymeyer *et al.*, 2000). Therefore, in addition, we evaluated ϕ-values from the thermodynamic perspective:

$$\phi = \frac{\Delta\Delta G^{\ddagger}}{\Delta\Delta G^0}, \tag{2}$$

where $\Delta\Delta G^{\ddagger}$ is the change in the free energy of the transition-state ensemble with the mutation present. The evaluation of Equation (2) combines the use of the multiple histogram method (Ferrenberg and Swendson, 1989; Kumar *et al.*, 1995) to estimate free energies, P_{fold} analysis (Du *et al.*, 1998) to obtain representative structures comprising the transition-state ensemble, and defining appropriate reference states for evaluating a meaningful $\Delta\Delta G^{\ddagger}$ and $\Delta\Delta G^0$, and is discussed more thoroughly in the methods section. The ability to directly probe the structure of the transition state allows us to bypass the hazardous interpretation of complicated kinetic analysis and to successfully predict ϕ-value trends for the five mutations from the thermodynamics. Furthermore, the level of resolution of the model allows us to directly predict what experiments seek in regard to protein engineering studies of protein folding—namely the residues or portions of the polypeptide chain that contribute to the bottleneck in the folding of the wild-type protein. We therefore conclude that minimalist models such as these provide a good level of resolution for answering broader questions concerning fold topology constraints on the transition state ensemble (Alm and Baker, 1999; Martinez and Serrano, 1999) and further validate the promise of using our minimalist model and design tools to tackle protein engineering objectives and proteomics endeavors that will play a significant role in the genomic-scale biology of the future.

Table 1. Bead and Dihedral Sequences for the Wild-Type Model[a]

Sequence	LBLBLBLBBN . NNLBBLBBBB . BNNNLLBLLB . BLLBNBLBLB
2° Structure	EEEEEETEH . THEEEEEEEE . HHEHHHHHHH . HHHEHTEEEE
Sequence	LBBNNNBBBL . BLBLBL
2° Structure	EEETTTEEEE . EEEE

[a]*B (hydrophobic), L (hydrophilic), N (neutral), E (extended), H (helix), T (turn).* The residues and 2 structure highlighted in gray denote sequence changes relative to the WT sequence reported in Sorenson (2000a).

RESULTS

A combination of positive and negative design techniques (Sorenson and Head-Gordon, 1999) were used to converge on an appropriate wild-type Protein L sequence for study. The "wild-type" sequence used in this work was a variant of the previously published sequence for a protein L model (Sorenson and Head-Gordon, 2000a). This new sequence, displayed in Table 1, differs from the old sequence in one bead flavor and three dihedral point mutations (highlighted in yellow) that were selected to further optimize the folding thermodynamics and kinetics. Our initial attempts to mutate the previously published sequence in order to compare it with experiment found that our simulated mutation data could not be easily explained, with mutations intended to destabilize the protein actually stabilizing it and vice versa. The sequence used here has been much more optimized and mutations away from it appear to introduce detrimental effects in every mutation examined, similar to what is observed in real protein mutagenesis experiments. In particular, a turn dihedral was added in the first beta strand in the newer sequence, because it shifts the native state to a lower energy native that was competing with the native state in the original sequence, and the additional mutations reduced the energy barriers and frustration of the landscape and increased the foldability of the model.

Table 2 summarizes the five mutant sequences studied here and intended for comparison with the large amount of experimental data on the effects of sequence mutations on the folding of protein L (Gu *et al*., 1997, 1999; Kim *et al*., 1998, 2000). Mutants $\beta 1$ and $\beta 2$ are similar to the Gly $\rightarrow$ Ala mutations in the β-hairpin #1 and β-hairpin #2, respectively, studied experimentally by Gu *et al.* (1997) and Kim *et al.* (2000). Mutants α and α^* are similar to the Glu $\rightarrow$ Ile (E32I) helix mutant examined experimentally by Kim *et al.* (1998, 2000) and Gu *et al.* (1999), where a surface polar residue in the α-helix has been mutated to a hydrophobic residue. The predictions of the many different secondary structure predictions programs do not expect this mutation to significantly disrupt the formation of the helix (Kneller *et al*., 1990; Rost, 1996), so the main effect of this mutation would be to possibly destabilize the native state and slow down the folding by favoring alternative hydrophobic cores. The last mutant is an additional mutation of mutant $\beta 1$, denoted $\beta 1^*$, by changing a neutral bead to a hydrophobic one in the first β-hairpin turn to further probe the role of nonnative hydrophobic interactions in the folding process, another possible consequence of the experimental Gly $\rightarrow$ Ala mutation in the turn region.

Table 2. Description of Mutant Sequences[a]

Name	*Mutation*	*Description*	*Experiment*
$\beta 1$	t10h	Disruptive to first β-hairpin turn region	G15A
$\beta 1^*$	t10h, N12B	Disruptive to first β-hairpin turn region	G15V
$\beta 2$	t45h	Disruptive to second β-hairpin turn region	G55A
α	L29B	Surface hydrophobic substitution in α-helix	E32I
α^*	L26B	Surface hydrophobic substitution in α-helix	E32I

[a]The corresponding experimental mutations explored by Baker and coworkers are listed in the last column.

Thermodynamics

The folding thermodynamics for the wild-type and all five mutant sequences was examined using Langevin dynamics simulation and the multiple multidimensional histogram method, as described in Methods. A summary of some thermodynamic signatures for each of the six sequences is given in Table 3. The folding of the wild-type sequence is very similar to the previously characterized protein L/G sequence in Sorenson and Head-Gordon (2000a). Because this new sequence was designed to be better than the previous sequence, the folding transition is slightly more cooperative and the onset of glasslike kinetics is delayed to lower temperature.

We have previously shown that the folding for this model is cooperative with a relatively sharp folding transition and a close coincidence of collapse and folding temperatures (Sorenson and Head-Gordon, 2000a). The folding exhibits multiple pathways with at least two clear ways for folding to proceed: a favored pathway involving formation of the N-terminal β-hairpin first, followed by formation of the C-terminal β-hairpin; and a second pathway, higher in energy, involving formation of secondary structure in the opposite order. The first pathway shows some evidence for a metastable intermediate along the pathway, while the second pathway has no such third state (see Fig. 4a below and corresponding discussion by Sorenson and Head-Gordon (2000a)).

The stability of the mutants relative to the wild-type sequence can be probed by examining the relative population of the native state versus temperature, $P_{nat}(T)$.

$$P_{nat}(T) = \frac{\sum_{E,\chi<\chi_{nat}} \Omega(E,\chi)e^{-E/T}}{\sum_{E,\chi} \Omega(E,\chi)e^{-E/T}} \tag{3}$$

The resulting curves, shown in Fig. 1, resemble protein denaturation curves and can be interpreted similarly. It is readily apparent that the wild-type sequence is more stable than the mutant sequences, with the highest folding temperature ($P_{nat}(T) = 0.5$) of $T_f = 0.45$ (see also Table 3). Evidence for a less cooperative folding transition is clearly seen in the profile of mutants α and α^* and to a lesser extent for mutants $\beta 2$ and $\beta 1^*$. These results are to be expected with the notion that the effect of the mutations examined here is to generally destabilize the native state by stabilizing nonnative structures, leading to lower folding temperatures and less cooperative folding. The details of how this destabilization is effected by each sequence can in turn be examined in their folding thermodynamics and the shape of their underlying free-energy landscape, and, in the following, we discuss this remaining thermodynamic data on a case-by-case basis for the five mutant sequences.

Mutants $\beta 1$ and $\beta 1^*$. Both mutants destabilize the formation of the first β-hairpin turn relative to the wild type by the mutation of the middle turn dihedral to a helix dihedral (Table 2). In the native-state structure, this dihedral has a value of $\psi = 320^\circ$, putting it close to a minimum in the turn potential, but in a local minimum for the helix potential (see Fig. 2). The native-state structure for these mutants is the

Table 3. Thermodynamic Parameters for the Wildtype and Mutant Sequences[a]

Sequence	T_f	gap/T_f	Z
wt	0.45	20.3	1.48
$\beta 1$	0.44	24.2	1.73
$\beta 1^*$	0.41	16.0	1.27
$\beta 2$	0.37	14.6	1.44
α	0.43	15.7	1.40
α^*	0.42	15.2	1.29

[a] T_f is the folding temperature and T_θ is the collapse temperature. The energy gap is defined as $\langle E_{nonnat}\rangle - \langle E_{nat}\rangle$ at the folding temperature. Z is the energy gap divided by the spread in energy of nonnative states $\langle \Delta E_{nonnat}\rangle$.

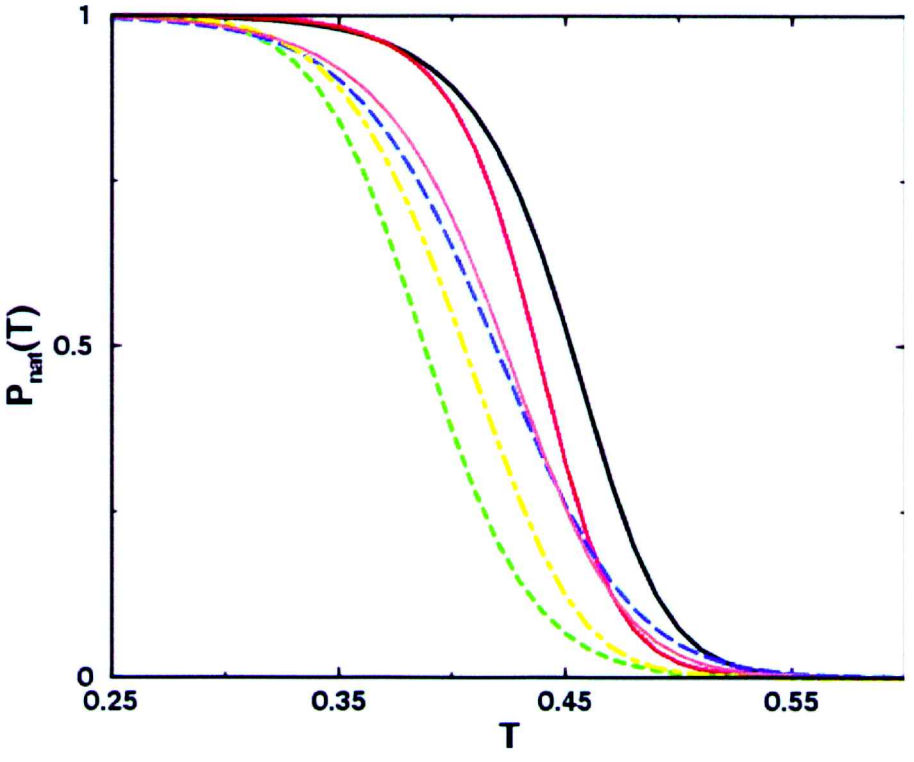

FIG. 1. Relative population of the native state versus temperature for the wildtype and mutant sequences. Legend: (from right to left) wt, $\beta 1$, α^*, α, $\beta 1^*$, $\beta 2$.

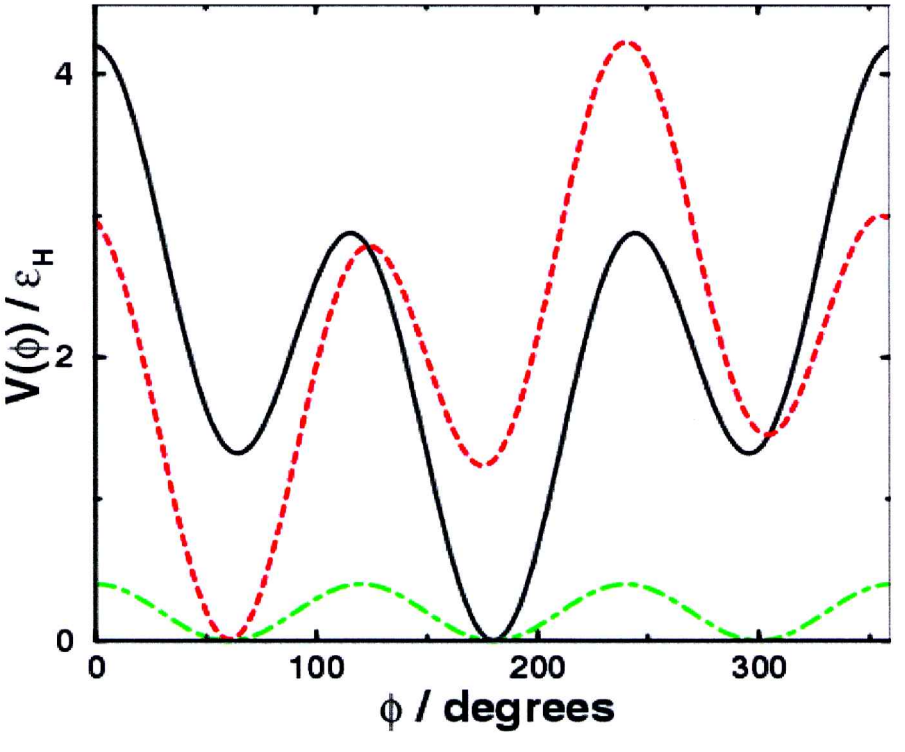

FIG. 2. Comparison of dihedral potentials. The helical potential has been shifted by a constant to have its global minimum at V = 0. Legend: extended (solid), helical (dashed), coil (dot-dashed).

same structure, with this dihedral now in a local minimum. This is similar to the effect of the experimental Gly → Ala mutation in the same region in protein L (Gu *et al.*, 1997; Kim *et al.*, 2000) that forces the turn residue to favor nonnative regions of the space of backbone dihedrals. Mutant $\beta 1^*$ further mutates the mutant $\beta 1$ sequence in the first turn region from an N residue to a B residue, to simulate the possible effect of nonnative hydrophobic contacts on the folding for this type of mutation.

A clear difference between the folding of these mutants and the wild-type sequence can be seen in Fig. 3, which shows the fluctuations of the radius of gyration, $\Delta^2 R_g(T) = \langle {R_g}^2(T) \rangle - \langle R_g(T) \rangle^2$, as a function of temperature. The plot measures the collapse transitions of the chain versus temperature, with peaks representing temperatures where the chain fluctuates highly between less collapsed and more collapsed structures. The collapse profile for mutants $\beta 1$ and $\beta 1^*$ shows a phase near T $\sim$ 0.47 not observed in the other sequences. Examination of the potentials of mean force for the secondary structure order parameters versus R_g (Sorenson and Head-Gordon, 2000c) shows that the large fluctuations are due to fluctuations about states where either β-hairpin #1 or β-hairpin #2 is formed with a partially formed helix. Unraveling of the formed β-hairpin #1 or the formation of the second hairpin leads to large changes in R_g. This particular phase does not appear in the wild-type sequence, presumably because the chain spends less time in a nonnative state for β-hairpin #1.

The detrimental consequences of nonnative structure formation for β-hairpin #1 is made more evident by looking at the free-energy landscape as a function of $\chi_{\beta 1}$ and $\chi_{\beta 2}$, the order parameters measuring formation of β-hairpins #1 and #2, respectively (Fig. 4). The asymmetry in the free-energy landscape, noted previously for a variation of the wild-type sequence (Sorenson and Head-Gordon, 2000a), is slightly changed for mutant $\beta 1$. While paths involving formation of either hairpin are still favored in the wild-type sequence, the mutant sequence shows evidence that it is now harder to follow the less dominant pathway, involving formation of β-hairpin #2 first followed by β-hairpin #1. While the two landscapes are mostly similar, the differences surrounding this second pathway are entirely reproducible with mutant $\beta 1^*$ (Sorenson and Head-Gordon, 2000c).

This result would appear at first to be counterintuitive, as the expected result of destabilizing the formation of the first β-hairpin would be to remove or destabilize the dominant pathway. That the local sequence mutation in the first β turn affects formation of the second β-hairpin, which is located 25 beads away in the sequence, indicates that the nonnative structures now being formed by the first β-hairpin act in such a way as to disfavor folding by the second pathway. This is presumably due to a topological restriction now introduced on the chain. When β-hairpin #2 forms with β-hairpin #1 in a nonnative structure, it looks to be sterically difficult for β-hairpin #1 to find the native state. The observation that a sequence mutation that is intended to have a local effect on structure can have very nonlocal consequences points to the difficulties in analyzing the effect of sequence mutations which are often conceptualized as acting locally.

Mutant $\beta 1^*$ exhibits an overall very similar folding pathway and free energy landscape as mutant $\beta 1$. From Fig. 2, we can see that the effect of the extra B bead is to further destabilize mutant $\beta 1$, although the cooperativity of folding is not nearly as affected as in the other hydrophobic mutation examined here, mutant α. As could be expected, the destabilizing effect of an L, $N \rightarrow B$ mutation is environment-dependent. The introduction of a B bead at the turn region in the sequence is not too detrimental because the remaining N beads surrounding this residue discourage nonnative B–B contacts. This effect will be discussed below for mutants $\beta 1$ and $\beta 1^*$.

Mutant $\beta 2$. The sequence for mutant $\beta 2$ incorporates the destabilizing mutation of a turn dihedral to a helix dihedral in the second β-turn. The resulting folding thermodynamics show the most destabilization of all of the sequences studied here (Table 3). The reason for the large native-state destabilization and consequently poor folding profile can be traced to another possible and likely common effect of sequence mutations. Simulated annealing of the sequence for mutant $\beta 2$ revealed that the new native-state structure in the second β-hairpin is slightly shifted with a new register of the β-strands in the hairpin being favored. The wild-type native-state structure is only slightly higher in energy however and now serves as a misfolded trap on the free-energy landscape, with a proven strong basin of attraction. Unlike the mutation in the first turn region, the collapse profile is very similar for this mutant compared to the wild type (Fig. 3). The combination of a lowered folding temperature without a concomitant change in the collapse temperature would also lead us to expect a less cooperative folding transition and slower kinetics (Klimov and Thirumalai, 1996, 1998).

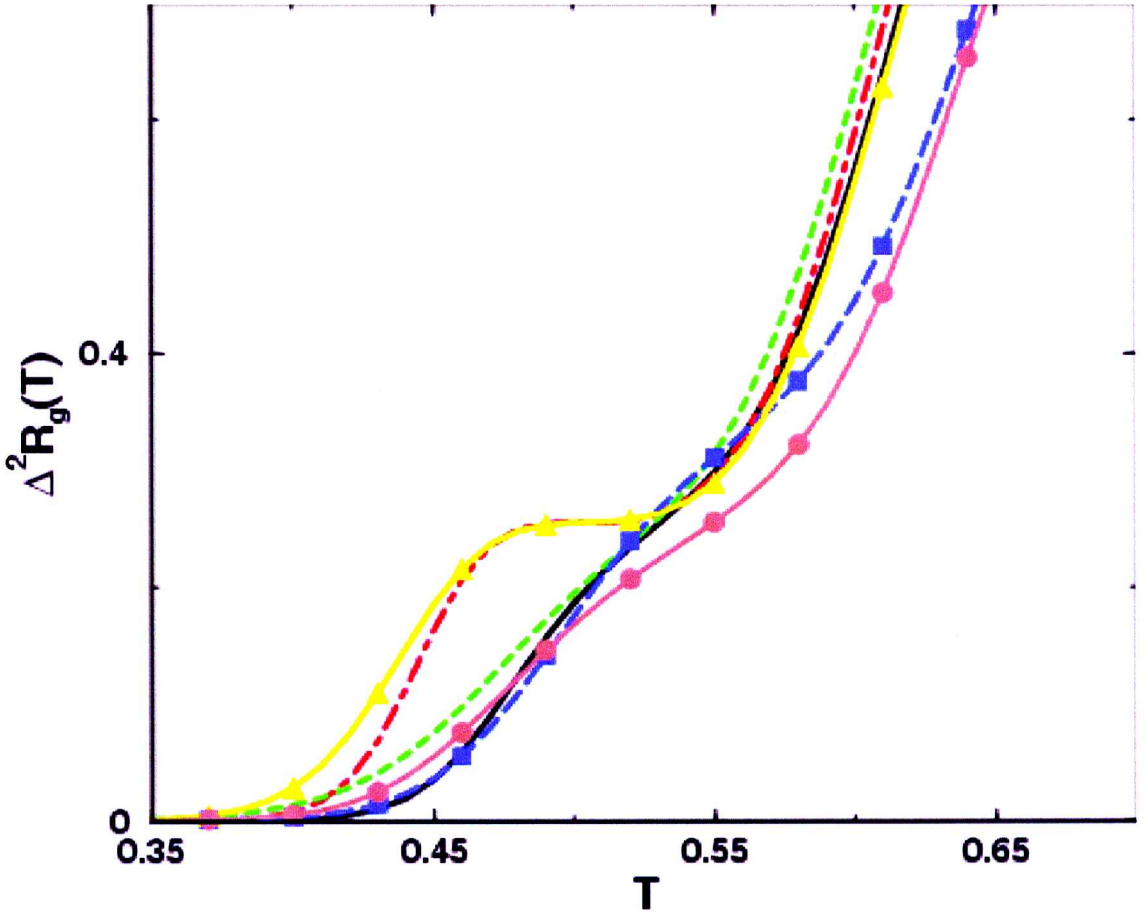

FIG. 3. Mean-square fluctuations in radius of gyration versus temperature for the wildtype and mutant sequences. Legend: wildtype (solid), $\beta 1$ (dot-dashed), $\beta 1^*$ (solid, triangles), $\beta 2$ (dashed), α (dashed, squares), α^* (solid, circles).

Mutants α and α^*. The sequences for mutants α and α^* were constructed to reproduce the destabilizing aspects of the E32I mutation experimentally studied in protein L (Kim *et al.*, 1998; Gu *et al.*, 1999). The experimental mutation involved changing a hydrophilic residue on the surface side of the amphiphilic α-helix to a hydrophobic residue. As stated previously, a number of secondary structure prediction servers show that this mutation should not disrupt the helix (Kneller *et al.*, 1990; Rost, 1996); however, a possible outcome of the mutation would be to slow down folding by confusing the search for the native hydrophobic core.

The combined thermodynamic evidence shows that this is precisely what happens for mutants α and α^* in our model. Figure 3 indicates that radius of gyration fluctuations are suppressed at higher temperatures relative to all of the other sequences. This shows that the chain is condensing to a more collapsed state at a higher temperature, and the ultimate collapse temperature is pushed higher. At the same time, the folding temperature is not raised (Fig. 2), so the kinetics and cooperativity are again expected to be poor (Klimov and Thirumalai, 1996, 1998). The increased number of alternative hydrophobic cores also leads to the noncooperativity exhibited in the shallower slope of the folding transition in $P_{nat}(T)$.

The best evidence that the helix mutation hampers the search for the native state can be seen in a plot of helix formation versus temperature for mutant α (Fig. 5). At higher temperatures, the other five sequences have very similar profiles for the formation of the helix as the temperature is decreased—as would be expected since they share identical sequences in the helix region. Mutant α, on the other hand, shows a pronounced delay in formation of the native helix as the temperature is lowered. The extra contacts now made by the helix inhibit the formation of native helical structure, and the loss of a clear amphiphilic helix

interferes with the proper alignment of secondary structure necessary for stabilizing the native helix. On this basis, we might expect mutant α^* to fare better; this is supported in the kinetic analysis below.

Kinetics

The kinetics of the folding process was tabulated for all six sequences at $T = 0.45$. This choice of temperature represents the folding temperature for the wild-type sequence and a temperature where all six sequences are expected to fold with exponential kinetics. This last condition aids in relating the observed folding kinetics to rate constants. The *in vitro* experiment corresponding to this study would be to compare the folding rate constants of various mutations, holding the temperature constant or maintaining the same denaturant concentration (Gu *et al.*, 1997; Khorasanizadeh, 1996).

The distribution of first-passage times for the β-turn mutants $\beta 1$, $\beta 1^*$, and $\beta 2$ are shown in Fig. 6 in comparison to the wild type. Similarly to previous work with the protein L/G model (Sorenson and Head-Gordon, 2000a), the folding at this temperature can be well characterized with two exponentials, representing folding pathways falling into two general classifications: a fast pathway and a slow pathway. The data for bi-exponential fits to the kinetic data are shown in Table 4. The quality of fit was measured by the χ^2 value:

$$\chi^2 = \frac{1}{N}\sum_i (y_i - y_i^0)^2 \tag{4}$$

where $\{y_i\}$ are the fit points and $\{y_i^0\}$ are the data points. Single and stretched exponential fits were also tried, but were much poorer than the bi-exponential fits for all six sequences.

As would be expected from the above analysis, mutant $\beta 2$ does fold slower than the wild-type sequence along the fast pathway, although the slow pathway has very similar kinetics at long times. This suggests the possibilty that mutations can nonuniformly affect the folding pathways, depending on which one is examined. Mutant $\beta 1$, in contrast, has extremely similar fast pathway kinetics, but experiences a slow-down along the slow pathway. As expected, mutant $\beta 1^*$ does not show very different kinetics from mutant $\beta 1$, although folding along the fast pathway is a bit slower, presumably due to nonnative B–B contacts.

The folding kinetics of mutants α and α^* are shown in Fig. 7. The significant misfolding experienced by mutant α suggested by $\chi_H(T)$ is very evident in this data as well. While the fast pathway looks to have a similar rate constant to the wild type, fewer chains are partitioned into this pathway, as evidenced by a lower value of a (the preexponential factor, see Table 4). The slower pathway, involving folding through misfolded intermediates, dominates the folding and has the slowest rate constant of the mutants examined here. Mutant α^* proves to be kinetically more robust, suggesting the importance of sequence context in determining the effects of surface hydrophobic substitutions.

Kinetic ϕ-values

The analysis of the change in protein kinetics and stability induced by site-directed mutations has been facilitated by the introduction of the ϕ-value (Equation (1) and in Matouschek *et al.* [1989]). Knowledge of the ϕ-value for a given mutation is intended to provide insight into the role of that residue in forming the transition state. If the residue participates in the structure of the transition-state ensemble, then we would expect a ϕ-value near 1. If instead the residue plays a relatively small role in stabilizing the transition state, then the ϕ-value should be closer to 0.

The calculation of ϕ-values provides a valuable method for comparison of experimental results with theoretical modeling of the same systems (Nymeyer *et al.*, 2000; Portman *et al.*, 1998). In practice, theoretical studies can attempt to use Equation (1), or alternatively they can calculate a similar value from the thermodynamic perspective using Equation (2). Recent work shows that the ϕ-values calculated from the thermodynamic approach are similar to the kinetic definition in Equation (1) only in the case that the energy landscape is not overly frustrated (Nymeyer *et al.*, 2000). The disadvantage of this approach lies in the additional complexity of properly identifying the transition-state ensemble, a nontrivial task as we move to more realistic protein models.

The interpretation of ϕ-values from experiment rests on several conditions. One necessary assumption is that the mutation causes a relatively small change in the kinetics and stability (Burton *et al.*, 1997). If this

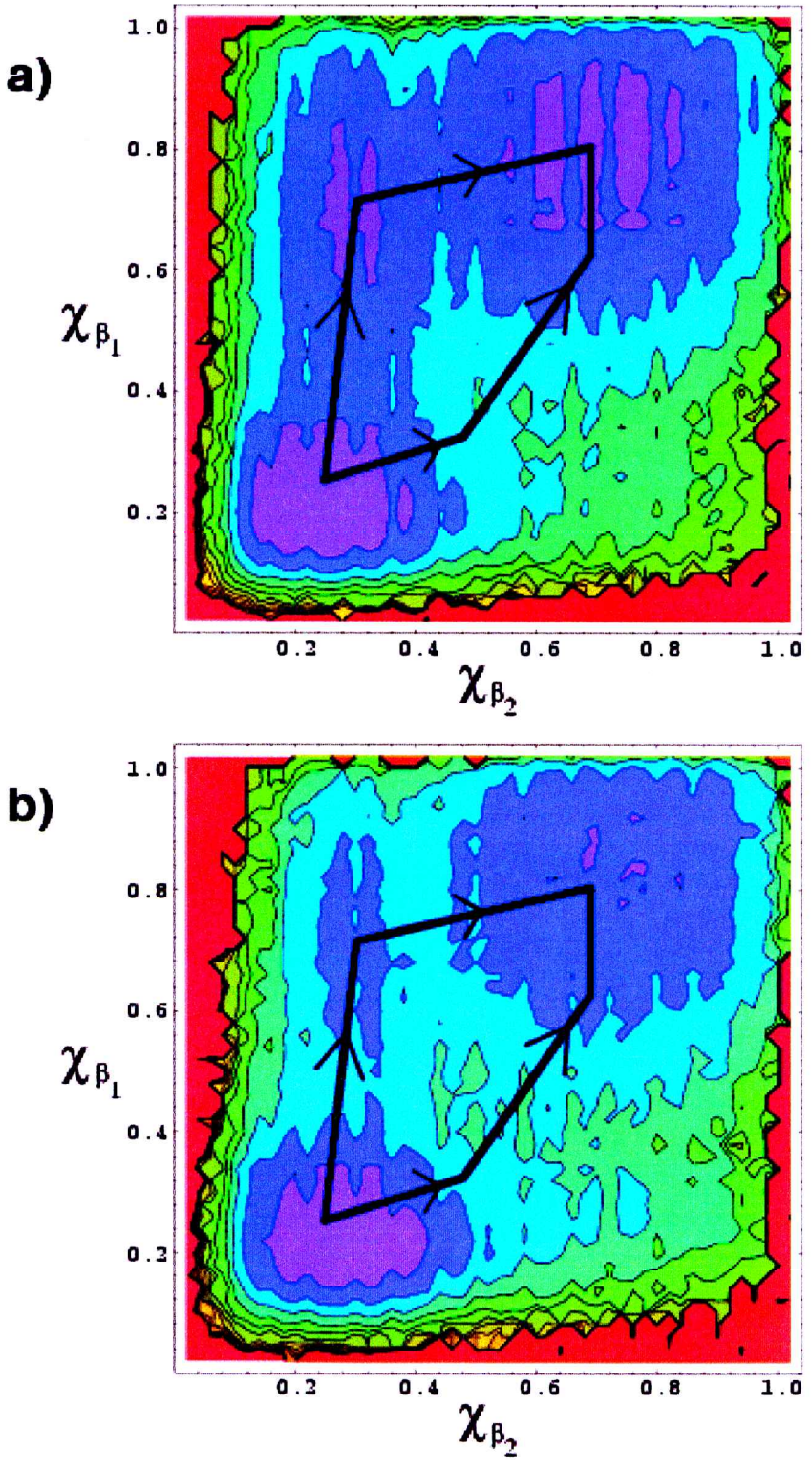

FIG. 4. (**a**) The potential of mean force as a function of $\chi_{\beta 1}$ and $\chi_{\beta 2}$ for the wildtype sequence at $T = 0.45$. (**b**) The potential of mean force as a function of $\chi_{\beta 1}$ and $\chi_{\beta 2}$ for mutant $\beta 1$ at $T = 0.43$. The pathways correspond to alternate folding pathways observed in simulation trajectories, with the dominant pathway being the one on the left.

is not the case, then the identification of the ϕ-value as a parameter describing local structure formation near the mutation site is invalid; instead, nonlocal effects distributed along the chain might be playing an essential role in the observed ϕ-value. A second assumption is that the value does not reflect structure formed in the unfolded state (Matsouschek *et al.*, 1989; Villegas *et al.*, 1998). If the unfolded basin of states contains significant structuring around the residue in question, then the measured ϕ-value will not necessarily correlate with transition-state structure formation.

The most crucial condition is that the folding of the protein chain is adequately described by a two-state picture with an unfolded basin of states, a folded basin of states, and a clear transition-state region separating these two basins. When this picture breaks down, as with current funnel-like conceptions of the free energy landscape for folding (Onuchic *et al.*, 1997; Dill and Chan, 1997), the simple interpretation of the ϕ-value is no longer straightforward (Nymeyer *et al.*, 2000; Burton *et al.*, 1997). A nice experimental test of this hypothesis is the comparison of ϕ-values derived from both unfolding and folding rates (Gu *et al.*, 1997); if the folding is primarily two-state, then the two values should complement each other. In addition to the possibility of multiple states along a single folding pathway, the presence of multiple folding pathways instead of a single dominant pathway will also greatly complicate the interpretation of ϕ-values. In all of these cases, there exist the possibilities of multiple transition-state ensembles all influencing the kinetics, and the structural effects of a single mutation might not be immediately clear from a single measured ϕ-value.

Various issues also arise in the calculation of ϕ-values from theory and simulations. Work with the present model (and similar models) has shown that collapsed states play an important role in the observed kinetics (Du *et al.*, 1998; Honeycutt and Thirumalai, 1990). In this case, the multistate nature of the folding process complicates the extraction of a single ϕ-value from the kinetics. In addition, multiple folding pathways have been observed to play a role in the kinetics of the present model (Sorenson and Head-Gordon, 2000a). One approach we have considered in this work has been to identify, where applicable, a fast rate constant and a slow rate constant for the folding process, representing the separation of folding scenarios into fast and slow pathways. From these rate constants and Equation (1), ϕ_{fast} and ϕ_{slow} can be identified ideally corresponding to the effects of the mutation on the transition states for the fast and slow pathways.

As mentioned above, unlike small protein sequences which often exhibit clear two-state single exponential kinetics, our model is best characterized as partitioning into fast and slow pathways, each governed by an activated process. A natural way then to compute ϕ-values is to isolate values for each pathway. Table 5 shows the result of applying Equation (1) to the rate constants corresponding to the kinetic fits in Table 4. The results are mixed; it appears that this approach tends to predict low ϕ-values. This might occur because our chosen mutations do not sufficiently perturb the kinetics, or it might reflect difficulties in extracting a single rate constant from the complex kinetics exhibited by the model.

The best agreement with the experimental ϕ-values can be found for mutant $\beta 2$. Experimental mutation of the second β-turn in protein L was found to destabilize the native state greatly, but to not as dramatically affect the folding kinetics (Gu *et al.*, 1997; Kim *et al.*, 2000), similar to our finding. Mutant α shows much higher ϕ-values for both pathways than the single experimental ϕ-value, but this can be easily understood in terms of nonnative hydrophobic contact formation and the stabilization of misfolded structures, seen in our model but not observed experimentally.

Several of the mutations explored here show the difficulties with calculating and interpreting ϕ-values when the underlying assumptions permitting their use are in question (Nymeyer *et al.*, 2000; Burton *et al.*, 1997). Mutants $\beta 1$ and $\beta 1^*$ are interesting in this respect because the most obvious change in their folding landscape is the perturbation of a pathway not connected to the local mutation. This violates the assumption of locally acting mutations. Similarily, mutant α also exhibits nonlocal effects upon mutation by bringing together portions of the chain that would not have previously been in contact. In all of these cases, the absence of single exponential kinetics also clearly makes the calculation of rigorous ϕ-values more difficult.

Thermodynamic ϕ-values

Given the problems inherent in calculating meaningful ϕ-values in systems that display strong bi-exponential kinetics such as our present model, we pursued the calculation of ϕ-values from the thermodynamic perspective (Equation (2)). This is possible in the present case through our use of the multiple-histogram method to estimate free energies and P_{fold} analysis to obtain representative structures comprising the transition-state ensemble (see the methods section).

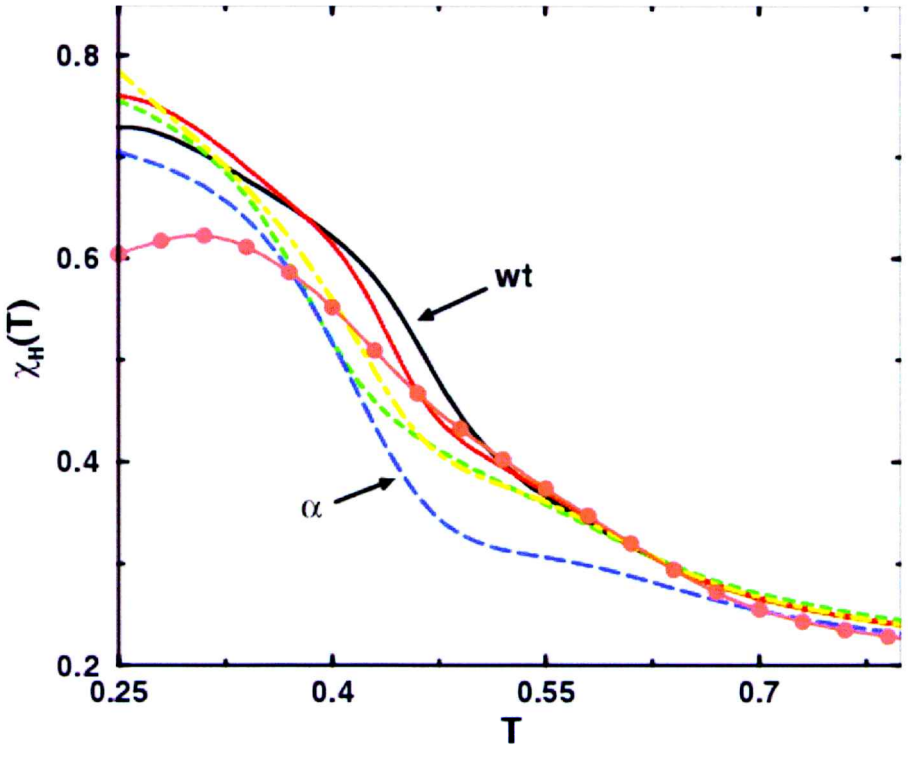

FIG. 5. Helix formation versus temperature for the wildtype and mutant sequences. Legend: (from right to left at $\chi_H(T) = 0.5$) wildtype, $\beta 1$, α^* (circles) $\beta 1^*$, $\beta 2$, α.

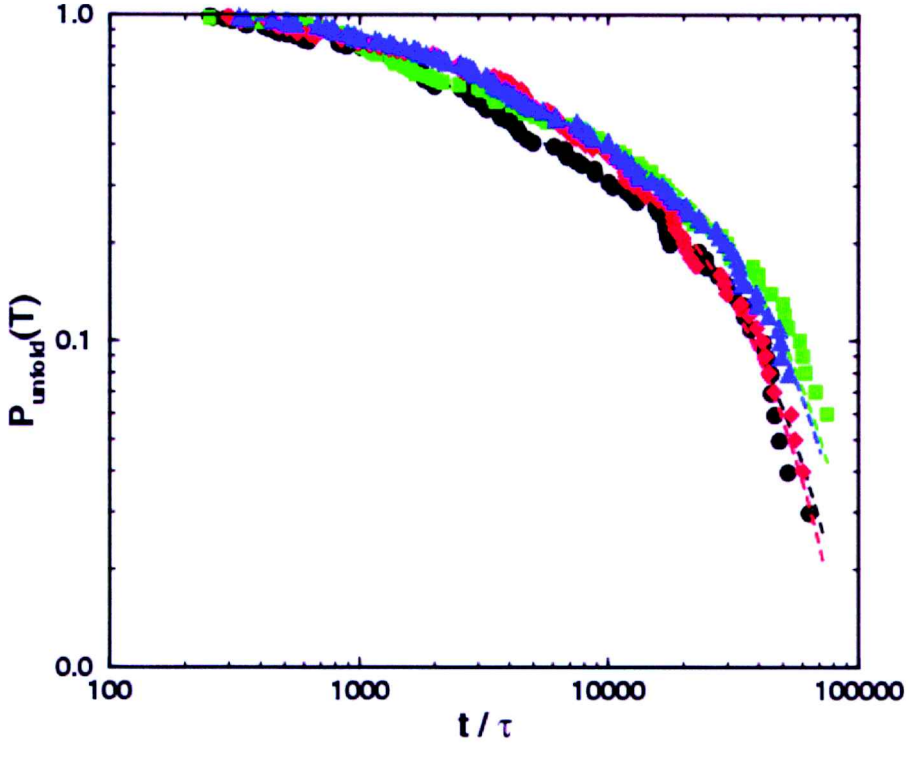

FIG. 6. Percentage of unfolded states versus time at $T = 0.45$ for wildtype (circles), mutant $\beta 1$ (squares), mutant $\beta 1^*$ (triangles), and mutant $\beta 2$ (diamonds).

Table 4. Parameters for Bi-exponential Fits to the Kinetic Traces in Figs. 6 and 7[a]

Sequence	a	τ_{fast}	$1-a$	τ_{slow}	$\chi^2/10^{-4}$
wt	0.52	1700	0.48	23000	4.3
$\beta 1$	0.45	1500	0.55	30000	3.4
$\beta 1^*$	0.47	2700	0.53	29000	6.4
$\beta 2$	0.37	2100	0.63	19000	2.6
α	0.34	2200	0.66	42000	1.0
α^*	0.65	3300	0.35	44000	0.8

[a]Fits are of the form $a\exp(-t/\tau_{fast}) + (1-a)\exp(-t/\tau_{slow})$.

Table 5. Kinetic ϕ-Values at $T = 0.45$ for Folding along the Fast and Slow Pathways for the Mutant Sequences

Sequence	$\Delta\Delta G^0$	$T \ln t_{fast}$	$T \ln t_{slow}$	ϕ_{fast}	ϕ_{slow}	$\phi_{experiment}$
$\beta 1$	0.37	−0.04	0.11	−0.1	0.3	0.8
$\beta 1^*$	0.93	0.21	0.09	0.2	0.1	0.7
$\beta 2$	1.24	0.09	−0.10	0.1	−0.1	0.2
α	0.38	0.12	0.27	0.3	0.7	0.06
α^*	0.53	0.30	0.29	0.6	0.5	0.06

Table 6. Thermodynamic ϕ-Values at $T = 0.45$ for the Mutant Sequences

Sequence	$\Delta\Delta G^0$	$\Delta\Delta G^{\ddagger}$	ϕ_{thermo}	$\phi_{experiment}$
$\beta 1$	0.37	0.86	2.3	0.8
$\beta 1^*$	0.93	1.2	1.2	0.7
$\beta 2$	1.24	0.68	0.5	0.2
α	0.38	0.13	0.3	0.06
α^*	0.53	−0.14	−0.3	0.06

Our approach was to calculate $\Delta G^{\ddagger}$ and ΔG^0 for each sequence by using an appropriate reference state common to each sequence. This reference state was chosen to be a set of 100 denatured chains drawn from a high-temperature simulation of the wild-type sequence at a temperature where the properties of all the models are sequence-independent (self-averaging). Calculating the free energy of this reference ensemble for each model then provides an appropriate reference state for measuring $\Delta G^{\ddagger}$ for each model such that $\Delta\Delta G^{\ddagger}$ can be meaningfully calculated. Similarly, $\Delta\Delta G^0$ is calculated by using a common reference state. In practice, these reference states take care of the issue that different sequences have different net values for their partition functions.

For each sequence, $\Delta G^{\ddagger}$ was calculated at $T = 0.45$ by finding a set of transition-state structures as described above and averaging the free energy $(-0.45 \ln P(R_g, \chi, \ldots)$, as calculated by the multiple histogram method) of the lowest ten structures and subtracting the free energy of the denatured ensemble to construct a free-energy difference. Also, ΔG^0 was straightforwardly calculated from $(-0.45 \ln P_{nat}/P_{non\text{-}nat})$, using the definition of native and nonnative states (see Equation (3)).

The ϕ-values that result from these quantities are shown in Table 6, and are seen to be consistent with the observation of the importance of the formation of the first β-hairpin in the transition-state structure. While we would be hard pressed to expect quantitative agreement with the experimental ϕ-values at this level of abstraction, the overall trends agree very well with the experimental findings for the folding of protein L ($r^2 = 0.84$). Unlike the problematic ϕ-values calculated from the kinetic analysis, these ϕ-values agree well with our intuition about the role of various elements in the folding process for this model

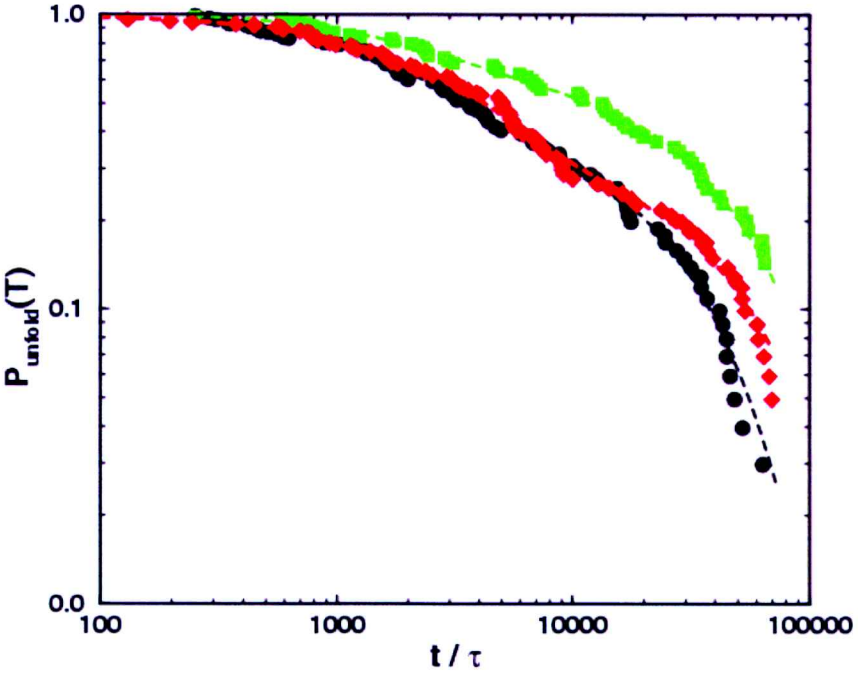

FIG. 7. Percentage of unfolded states versus time at $T = 0.45$ for wildtype (circles), mutant α (squares), and mutant α^* (diamonds).

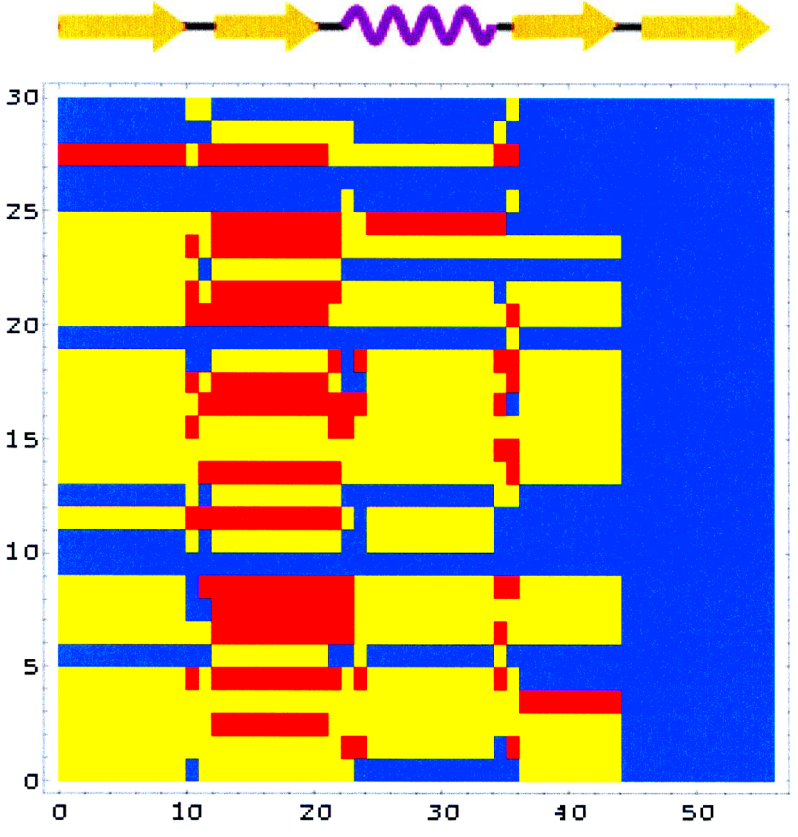

FIG. 8. Location of native-like structure formation (χ) for thirty low free-energy transition states ($0.45 \leq P_{fold} \leq 0.55$) for the wildtype sequence. The thirty structures are ordered with respect to increasing free energy, with the structure with the lowest free energy at the bottom of the plot. Legend: most native-like structure (red), some native-like structure (yellow), no native-like structure (blue).

(Sorenson and Head-Gordon, 2000a). Our model appears to more closely mimic the folding of protein L versus protein G (McCallister *et al.*, 2000), a point that is discussed further below.

The disagreement between ϕ-values calculated from the kinetic and thermodynamic pictures points to several places where the folding of the model differs from experiment. As emphasized by Nymeyer *et al.* (2000), the degree of frustration in a protein folding model can strongly influence the agreement of ϕ-values found from both methods. The disagreement in our case might represent a relatively large amount of frustration in the folding of our model, as also evidenced by the heavy role of compact intermediates in the folding process. However, the more problematic issues for kinetic ϕ-value analysis in the context of this model are the difficulties of calculating a single rate constant for a clearly bi-exponential process, the complex role of multiple pathways, and the differing relative populations of slow and fast folders across different sequences. All of these issues break down the conventional picture underlying the interpretation of experimental ϕ-values and strongly discourage the use of kinetic ϕ-values in the present case.

In the final analysis, the ability to exactly probe the structure of the transition-state ensemble through P_{fold} analysis allows us to bypass the hazardous interpretation of kinetic ϕ-values. We can exactly know the structure of the transition-state ensemble without deducing this information through complicated kinetic analysis, and we can verify the changes in the transition state ensemble through the evaluation of thermodynamic ϕ-values.

The wild-type transition-state ensemble

Using the P_{fold} analysis described in Methods, we can extract structures corresponding to transition states for folding. The resulting structures comprising the low free-energy transition-state ensemble for the wild-type sequence are characterized in Fig. 8. A total of 105 transition-state structures were identified by P_{fold} analysis; only the thirty structures lowest in free energy, as determined by the multiple histogram method, are depicted in Fig. 8. The structural analysis summarized in Fig. 8 confirms our earlier identification (Sorenson and Head-Gordon, 2000a) of the importance of structure formation in the first part of the chain. We had previously identified the formation of the first β-hairpin to be more favorable on the basis of certain order parameter projections onto the free-energy surface (Sorenson and Head-Gordon, 2000a); the existence of many low free-energy transition-state structures along this pathway confirms the free-energy analysis.

The most nativelike structure appears to occur first in the second strand of the first β-hairpin. This is consistent with the observation that the most stabilizing hydrophobic contacts in the native-state structure occur using beads from β-strand #2 (Sorenson and Head-Gordon, 2000a); the proper formation of structure in this strand might serve to stabilize nucleation of the helix and first β-hairpin, which appear to be important structural elements of the transition-state ensemble. Experimental mutations that probed the region between the first β-hairpin and the helix showed intermediate ϕ-values and indicate that this region of the hydrophobic core is at least partially structured in the transition state (Kim *et al.*, 2000).

The diversity of structure in the transition-state ensemble is an important element that should be noted. While most of the low free-energy structures favor similar structure formation, several structures contain notably no nativelike structure (on this scale) and several structures contain a distinctly different pattern of structure formation. This further indicates the strong role of multiple pathways in the folding of this model and greatly complicates analysis of the kinetic effects of sequence mutations over a simple one-pathway picture. While a particular mutation might destabilize one pathway, the existence of low-lying alternate folding pathways not requiring structure formation in the same region of the chain makes the overall kinetic effect difficult to predict.

Comparing proteins L and G, we readily note the similarity in the relative importance of the first β-hairpin in the transition-state ensemble (Gu *et al.*, 1997). Interestingly, the folding of protein G has been long suspected to have a different kinetic folding mechanism than does protein L (Ramírez-Alvarado *et al.*, 1997; Blanco and Serrano, 1995; Park *et al.*, 1997) and recent protein engineering experiments have supported this (McCallister *et al.*, 2000), demonstrating that the second β-hairpin in protein G is likely more structured in the transition state. A priori, we would not have known whether our model would fold more like protein L or like protein G. However, as our model shares an identical topology with both proteins, we might expect aspects of each in the observed folding. Indeed, transition-state structures higher in free energy than the structures shown in Fig. 8 do show evidence for nativelike structure formation in the second half of the chain instead of the first (Sorenson and Head-Gordon, 2000a). The existence of

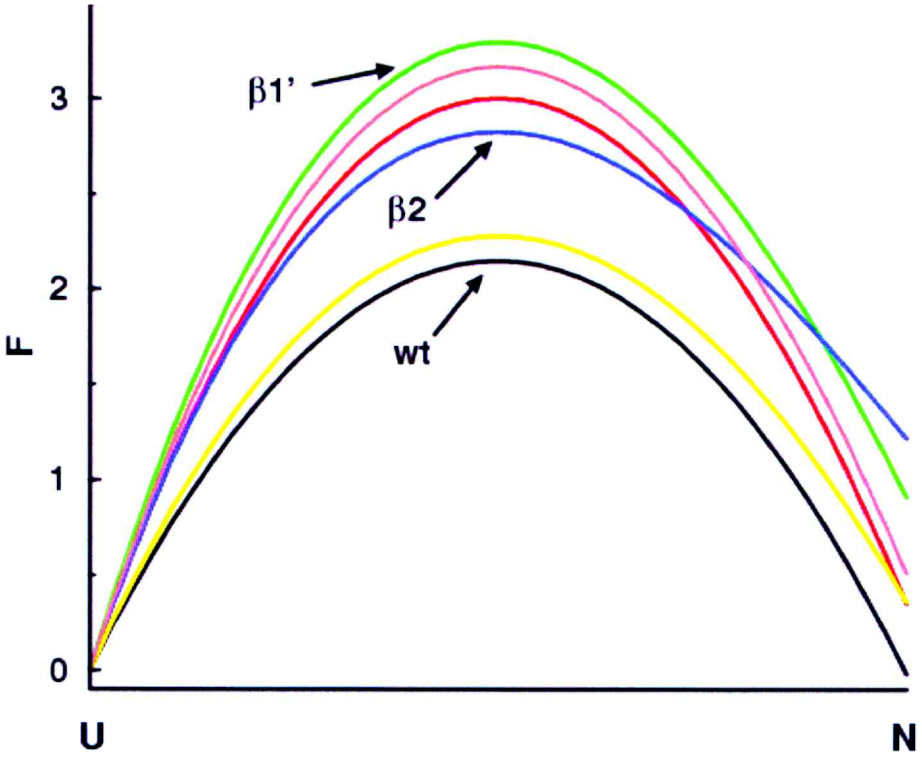

FIG. 9. Reaction coordinate plot at $T = 0.45$. The free energy differences between the unfolded and native states, and the "transition-state" barrier heights ($\Delta\Delta G^{\ddagger}$) as determined from the simulations (in units of ε_H), the multiple histogram method, and P_{fold} analysis described in Section 2. Legend: (from bottom to top) wildtype, α, $\beta 2$, $\beta 1$, α^*, $\beta 1^*$.

these transition-state structures confirms the overall similarity of our model to proteins L and G and further illustrates the multiple-pathway difficulties which might play a role in the interpretation of experimental protein engineering studies. In any case, we have provided a direct assessment of the nature of the transition state ensemble for wild-type protein, information that is only indirectly inferred by experimental mutation studies.

DISCUSSION AND CONCLUSIONS

Experimental comparison

A graphical summary of the simulated thermodynamic data for all of the sequences is presented in Fig. 9 and allows comparison to the many similar diagrams drawn in the analysis of experimental mutation studies. Each mutation was designed to allow comparison of folding thermodynamics and kinetics between simulation and experimentally characterized single-site mutants of protein L. The destabilized first-turn mutants β and $\beta 1^*$ did show moderate destabilization and somewhat slower kinetics, but not as much as to allow quantitative comparison to the corresponding experimental mutation, G15A, that exhibited a greater destabilized native state and significantly more perturbed kinetics (Gu *et al.*, 1997; Kim *et al.*, 2000). However, the modeled mutation does support a very similar asymmetry in the folding of the β-hairpins and the precedence of the first half of the chain of protein L involving the the first two strands and the helix (Gu *et al.*, 1997; Kim *et al.*, 2000). For example, the experimental study providing a more exhaustive ϕ-analysis for point mutations showed that the region between the first β-hairpin and the helix showed intermediate ϕ-values, indicating that this region of the hydrophobic core is at least partially structured in the transition state (Kim *et al.*, 2000). This is also present in our model as is evident when considering Fig. 8.

Gu *et al.* (1997) also studied a G15V mutant to verify that the observed result was due to a change in turn propensity rather than the formation of nonnative hydrophobic contacts. Our mutant $\beta 1^*$ showed some detrimental nonnative contact formation, but not a significant amount, especially when compared with the similar mutant α. This is consistent with the experimental result and suggests that the hydrophilic turn region in both the model and experiment is robust to a hydrophobic substitution. The surrounding hydrophilic groups are sufficient to deter the formation of strong nonnative hydrophobic contacts.

Mutant $\beta 2$ is an interesting case, where the introduction of a sequence mutation sufficiently destabilizes the native state to the point that a new native-state structure is favored. Because the two structures are still close to each other in energy and both possess well-designed funnels, they compete with each other in the folding process and cause the large destabilization seen here. The resulting thermodynamics and kinetics are similar to the experimental G55A mutant (Gu *et al.*, 1997), where it appears that the mutation affects the stability of the native structure much more than the folding kinetics. Whether or not the experimentally measured destabilization is due to formation of a slightly different native structure could be difficult to tell, as most experiments necessarily employ coarse-grained measures of when nativelike structure is obtained. The adoption of a slightly different native state is a distinct possibility for many experimental mutations and can cloud the analysis of the resulting stability and kinetics; structure determination for each mutant sequence would avoid this, at the expense of significantly more laboratory work.

The destabilizing effects of a hydrophobic substitution on the protein surface are evident in mutants α and α^* For mutant α, the chosen bead on the helix, #29, is not in an overly hydrophilic region of the sequence, and this fact, combined with the central location of the mutation, stabilizes misfolds and leads to noncooperative folding and much slower folding kinetics. The site mutation for mutant α^* is in a more hydrophilic region, and the resulting kinetic effect appears to be less severe. The corresponding helix mutation, E32I, in protein L does not show a change in folding kinetics from the wildtype (Kim *et al.*, 1998), although the rate constant for unfolding is increased, similarly to that seen here. This compares with the experimental observation for another region of the protein L sequence that surface hydrophobic substitutions do not overly confuse the search for the native state (Gu *et al.*, 1999). However, this work does indicate that the robustness of the sequence to hydrophobic substitution is somewhat dependent on the location of the chosen site, with certain combinations of substitutions being very unfavorable. We find a similar result comparing mutants α, α^*, and $\beta 1^*$, which all have a surface hydrophobic substitution but do not show equally perturbed kinetics. West and Hecht (1995) observe that amphiphilic helices across a large range of structures show very strong hydrophobic–hydrophilic patterning, suggesting that tampering with these patterns could lead to our observed result.

One issue, related to the level of abstraction of the model, is the similarity of the model to both protein L and protein G—proteins with nearly identical structures but less than 30% sequence homology and different folding characteristics (Gu *et al.*, 1997; Park *et al.*, 1997; Plaxco *et al.*, 1999), and secondary structure propensities (Ramírez-Alvarado *et al.*, 1997; Blanco and Serrano, 1995). As had been noted previously in Sorenson and Head-Gordon (2000a), the folding of our model shares characteristics with both sequences, so a necessary issue with comparison to a specific protein sequence is how closely our sequence models protein L and not a generic sequence for this structure. Overall, we find that the wild type sequence more closely resembles protein L than it does protein G. An interesting future study will perturb the wild type sequence in order to lower the free energy of the transition state ensemble that favors formation of the second β-hairpin relative to β-hairpin #1, to better understand the relative role of sequence details versus fold topology.

The combined evidence from multiple uses of models of this type indicate that nonnative collapsed states might play too large of a role in the folding kinetics when compared to real proteins. While the ability to calculate thermodynamic ϕ-values allows us to best confirm the model's underlying similarity to experiment, the difficulty in obtaining similar kinetic ϕ-values and the relatively small kinetic perturbation seen in Figs. 6 and 7 suggests areas where the model can be productively enhanced to better match the experimental picture.

Part of the reason for lack of quantitative agreement with experiment is that real proteins have structural features not modeled here that might better enable specific collapse. These include variation in attractive interactions (polar–polar, as well as hydrophobic), more cooperative formation of α-helices and β-sheets through specific backbone hydrogen bonding, and the cooperativity and uniqueness provided by the very

specific side-chain packing of the native-state core. Since our goal is to suppress unnecessary computational complexity while maintaining faithfulness to the experimental observables, we can suggest fruitful areas of improvement to the model that allow us to maintain this balance.

One area is in the construction of the dihedral potentials and mutations of the dihedral sequence. The parameter choices for the dihedral potentials are inspired by the earlier work pioneered by Thirumalai and coworkers (Honeycutt and Thirumalai, 1990; Guo and Thirumalai, 1994), and the effects of modifying these potentials has not yet been examined in this model. In this context, it would be important to explore the role of the relative barrier heights and minima in Fig. 1 in governing the observed changes when a mutation is made. Another possible modification of this model which preserves its essential simplicity but might add more modeling flexibility would be to increase the number of "flavors" in the model—i.e., allowing more bead types for a wider range of amino acid categories, changing the complexity of the interaction such as was done by Sorenson and Head-Gordon (1998), and/or incorporating a larger variety of dihedral potentials.

The goal of protein-engineering experiments that measure the mutant and wild-type folding rates is to (indirectly) correlate the role of that residue in forming the transition state. In principle, a model that is complex enough to reproduce important experimental trends and observables, but is simple enough to fully characterize, can provide information about the wild-type transition-state structure directly. A further objective of protein folding experiments is to address the relative role of the native-state topology versus the details of the sequence in order to capture the level of resolution necessary for protein folding prediction (Alm and Baker, 1999; Martinez and Serrano, 1999). Overall, the agreement of our minimalist model with experimental trends is excellent. While further improvements in our minimalist model will be pursued to realize better quantitative agreement with experiment in the future, we believe that the level of resolution required for robust folding predictions is obtainable with protein-folding models and design tools such as those described here.

METHODS AND MODELS

The energy function

The level of description of the protein chain and choice of parameters for the energy function have been extensively described in previous publications (Sorenson and Head-Gordon, 2000a, 2000b; Honeycutt and Thirumalai, 1994; Guo and Thirumalai, 1994). The essential details will be described here. The protein chain is modeled as a chain of beads of three flavors: hydrophobic (B), hydrophilic (L), or neutral (N). Attraction between the hydrophobic beads provides the energetic driving force for formation of a strong core, repulsion between the hydrophilic beads and other beads are used to balance the forces and bias the correct native fold, and the neutral beads serve as soft spheres with little repulsion and typically signal the turn regions in the sequence.

The potential energy for the model is

$$\begin{aligned} E = & \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 \\ & + \sum_{dihedrals} \left\{ A[1 + \cos\psi] + B[1 - \cos\psi] + C[1 + \cos 3\psi] + D\left[1 + \cos\left(\psi + \frac{\pi}{4}\right)\right]\right\} \\ & + \sum_{i,j \geq i+3} 4\varepsilon_H S_1 \left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - S_2 \left(\frac{\sigma}{r_{ij}}\right)^{6}\right]. \end{aligned} \tag{5}$$

Bond lengths are held rigid, and the bond angles are maintained by a harmonic potential with force constant $k_\theta = 20\varepsilon_H/(\text{rad})^2$ and equilibrium bond angle $\theta_0 = 105^\circ$. The basic energy unit, ε_H, corresponds to the minimum in the attraction energy between two B beads. The nonlocal interactions are given by $S_1 = S_2 = 1$ for BB interactions, $S_1 = 1/3$ and $S_2 = -1$ for LL and LB interactions, and $S_1 = 1$ and $S_2 = 0$ for all interactions involving N residues.

The dihedral potentials in Equation (5) are designed to simulate extended, helical, or turn regions of the protein sequence (Honeycutt and Thirumalai, 1990; Guo and Thirumalai, 1994). To specify a particular protein sequence in this model, the bead sequence and the dihedral sequence must both be specified. The parameter choices corresponding to the three possible dihedral states are A = 0, B = C = D = $1.2\varepsilon_H$ for helical, (A = $0.9\varepsilon_H$, B = D = 0, C = $1.2\varepsilon_H$) for extended, or (A = B = D = 0, C = $1.2\varepsilon_H$) for turn, and the resulting potential energy as a function of dihedral angle ψ are shown in Fig. 2. The advantages and limitations of this description have been addressed before by Sorenson and Head-Gordon (2000a). The dihedral potentials serve as potentials of mean force meant to reproduce the intrinsic secondary structure propensity of certain sequences of amino acids. Such potentials are required since important determinants of secondary structure in real proteins, such as backbone hydrogen bonding and side-chain steric restrictions, are "integrated" out at our level of description.

Simulation methods

Sequences were simulated with Langevin dynamics at constant temperature. The bond lengths were held rigid by solving the constraint equations of motion with the RATTLE algorithm. All further details about parameter choices and simulation methods have been presented several times before (Sorenson and Head-Gordon, 1999, 2000a, 2002). We use a simulated annealing protocol to find the global minimum for sequences in this model. The simulations are performed in reduced units, with the units of mass m, length σ, energy ε_H, and k_B all set equal to one; temperature is in units of ε_H/k_BT. The unit of reduced time is $\tau = \sqrt{m\sigma^2/\varepsilon_H}$.

Details of the folding thermodynamics for each sequence were found using the multiple multi-dimensional histogram method (Ferrenberg and Swendsen, 1989; Kumar *et al.*, 1995) and sampling over a range of temperatures. In this work, we collected six-dimensional histograms over energy and five order parameters—radius of gyration (R_g), native-state similarity (χ), native-state helix formation (χ_α), native-state β-hairpin #1 formation ($\chi_{\beta 1}$), and native-state β-hairpin #2 formation ($\chi_{\beta 2}$); χ is the order parameter for folding to the native state:

$$\chi = \frac{1}{M} \sum_{i,j \geq i+4}^{K} \theta(\varepsilon - |r_{ij} - r_{ij}^{nat}|) \tag{6}$$

where the double sum is over beads on the chain, r_{ij} and r_{ij}^{nat} are the distances between beads i and j in the state for comparison and the native state, respectively, θ is the Heaviside step function, and $\varepsilon = 0.2$ accounts for small fluctuations away from the native state structure. M is a normalizing factor to ensure that $\chi = 1$ when the chain is identical to the native state and $\chi = 0$ when the chain is in a random coil state. We note that the collection of histogram data over six dimensions requires a modification of the traditional array structure used for one-dimensional or two-dimensional histograms. Such data structures rapidly exhaust system memory when the number of dimensions increases. We avoid this problem in our work by using a hash table with linked lists for collision resolution (Cormen *et al.*, 1990).

The folding kinetics were probed by calculating first passage times for when the chain first enters the native-state basin of attraction. For the purpose of this study, this basin of attraction corresponds to $\chi > 0.42$; this choice is further justified by Sorenson and Head-Gordon (2000a).

P_{fold} Analysis. For the kinetic behavior of complex systems, characterized by many possible variables, it is often difficult to assign reaction coordinates and local transition-state structures along these coordinates. The identification of structures corresponding to a maximum along the free-energy profile of one order parameter does not necessitate that this subset of structures comprises the transition-state ensemble (Geissler *et al.*, 1999). Therefore, the problem in using Equation (2) resides in the difficulty of assigning reaction coordinates and locating transition-state structures along these coordinates. To more accurately characterize the proper transition-state ensemble, we employed a method proposed by Du *et al.* (1998). The method assigns a value, P_{fold}, to a particular structure corresponding to the probability that that particular structure will first fold to the native state before unfolding. Structures with P_{fold} values equal to 0.5 correspond to the transition-state ensemble for the model (Du *et al.*, 1998). The procedure is straightforward to computationally implement, although time intensive (Du *et al.*, 1998; Geissler *et al.*, 1999).

To apply this method to our current model of five mutants, we first sampled structures from our simulations corresponding to putative transition-state structures. "Putative" transition-state structures were originally isolated by requiring various combinations of χ_H, $\chi_{\beta 1}$, $\chi_{\beta 2}$ to correspond to their maximum free-energy values in a one-dimensional projection of free energy against these order parameters. Other putative structures were sampled by fixing χ at the maximum in free energy in $F(\chi)$. From this procedure, an ensemble of structures with $0.45 \leq P_{fold} \leq 0.55$ was isolated. These were further screened by selecting structures from independent simulation trajectories to avoid correlated sampling and restricting the set to structures with the lowest free energies as predicted by the multiple-histogram method.

To characterize the native-state similarity of these transition-state structures, the standard χ native-state similarity measure (Sorenson and Head-Gordon, 2000a) was modified to give structural information on a bead-by-bead level. We can write χ as the sum of pairwise order parameters, χ_{ij}, which are 1 when the distance between beads i and j is within 0.2σ of the corresponding distance in the native-state structure and 0 otherwise. Then we can define a local χ_i for each bead i as

$$\chi_i = \frac{1}{M}\left(\sum_{j=1}^{i-4} \chi_{ij} + \sum_{j=i+4}^{N} \chi_{ij}\right) \tag{7}$$

where the sums are over all of the beads in the chain except nearby beads and M is a normalization constant which ensures that χ_i is normalized properly for beads at the end of the chain (which have shorter bead–bead distances). Using this definition, we can see that native-state structures with $\chi = 1$ will also have $\chi_i = 1$ for each bead. On the other hand, for chains that are not identical to the native state, this local measure of native-state similarity identifies which segments of the chain are most nativelike. Having isolated what we believe is the transition-state ensemble for the mutant in question, we can evaluate the quantity $\Delta\Delta G^{\ddagger}$ in Equation (2) in order to calculate a ϕ-value from the thermodynamics to validate our model against ϕ-values from experiment.

ACKNOWLEDGMENTS

T.H.G. would like to acknowledge financial support from the U.S. Department of Energy Contract #DEAC-03-76SFOO098, and the University of California Berkeley for start-up funds. J.M.S. thanks the National Science Foundation for a Graduate Research Fellowship from 1997–2000.

REFERENCES

Alm, E., and Baker, D. 1999. Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.* 2, 189–196.

Blanco, F.J., and Serrano, L. 1995. Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.* 230, 634–649.

Burton, R.E., Huang, G.S., Daugherty, M.A., Calderone, T.L., and Oas, T.G. 1997. The energy landscape of a fast-folding protein mapped by Ala-Gly substitutions. *Nat. Struct. Biol.* 4, 305–310.

Cormen, T.H., Leiserson, C.E., and Rivest, R.L. 1990. *Introduction to Algorithms.* MIT Press, Cambridge, MA.

Dill, K.A., and Chan, H.S. 1997. From Levinthal to pathways and funnels. *Nat. Struct. Biol.* 4, 10–19.

Du, R., Pande, V.S., Yu, A., Grosberg, T., Tanaka, T., and Shakhnovich, E.I. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108, 334.

Ferrenberg, A.M., and Swendsen, R.H. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63, 1195–1198.

Geissler, P.L., Dellago, C., and Chandler, D. 1999. Kinetic pathways of ion pair dissociation in water. *J. Phys. Chem. B* 103, 3706.

Gu, H., Doshi, N., Kim, D.E., Simons, K.T., Santiago, J.V., Nauli, S., and Baker, D. 1999. Robustness of protein folding kinetics to surface hydrophobic substitutions. *Protein Sci.* 8, 2734–2741.

Guo, Z., and Thirumalai, D. 1994. Kinetics of protein folding: Nucleation mechanism, time scales, and pathways. *Biopolymers* 36, 83–102.

Gu, H., Kim, D., and Baker, D. 1997. Contrasting roles for symmetrically disposed β–turns in the folding of a small protein. *J. Mol. Biol.* 274, 588–596.

Honeycutt, J.D., and Thirumalai, D. 1990. Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci.* 87, 3526–3529.

Kim, D. E., Fisher, C., and Baker, D. 2000. A breakdown of symmetry in the folding transition state of Protein L. *J. Mol. Biol.* 298, 971–984.

Kim, D.E., Yi, Q., Gladwin, S.T., Goldberg, J.M., and Baker, D. 1998. The single helix in protein L is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.* 284, 807–815.

Khorasanizadeh, S., Peters, I.D., and Roder, H. 1996. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nat. Struct. Biol.* 3, 193–205.

Klimov, D.K., and Thirumalai, D. 1996. Criterion that determines the foldability of proteins. *Phys. Rev. Lett.* 76, 4070–4073.

Klimov, D.K., and Thirumalai, D. 1998. Cooperativity in protein folding: From lattice models with sidechains to real proteins. *Fold. Design* 3, 127–139.

Kneller, D.G., Cohen, F.E., and Langridge, R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 214, 171–182.

Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., and Kollman, P.A. 1995. Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comp. Chem.* 16, 1339–1350.

Martinez, J.C., and Serrano, L. 1999. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* 6, 1010–1016.

Matouschek, A., Kellis, J.T., Serrano, L., and Fersht, A.R. 1989. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* 340, 122–126.

McCallister, E.L., Alm, E., and Baker, D. 2000. Critical role of β–hairpin formation in protein G folding . *Nat. Struct. Biol.* 7, 669–673.

Nymeyer, H., Socci, N.D., and Onuchic, J.N. 2000. Landscape approaches for determining the ensemble of folding transition states: Success and failure hinge on the degree of frustration. *Proc. Natl. Acad. Sci.* 97, 634–639.

Onuchic, J.N., Luthey-Schulten, Z., and Wolynes, P.G. 1997. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* 48, 545–600.

Park, S-H., O'Neil, K.T., and Roder, H. 1997. An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core. *Biochemistry* 36, 14277–14283.

Plaxco, K.W., Millett, I.S., Segel, D.J., Doniach, S., and Baker, D. 1999. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol.* 6, 554–556.

Portman, J.J., Takada, S., and Wolynes, P.G. 1998. Variational theory for site-resolved protein folding free energy surfaces. *Phys. Rev. Lett.* 81, 5237–5240.

Ramírez-Alvarado, M., Serrano, L., and Blanco, F.J. 1997. Conformational analysis of peptides corresponding to all the secondary structure elements of protein L B1 domain: Secondary structure propensities are not conserved in proteins with the same fold. *Protein Sci.* 6, 162–174.

Rost, B. 1996. PHD: Predicting one-dimensional protein structure by profile based neural networks. *Methods Enzym.* 266, 525–539.

Sorenson, J.M., and Head-Gordon, T. 1998. The importance of hydration for the kinetics and thermodynamics of protein folding: Simplified lattice models. *Fold. Design* 3, 523–534.

Sorenson, J.M., and Head-Gordon, T. 1999. Redesigning the hydrophobic core of a model β–sheet protein: Destabilizing traps through a threading approach. *Proteins: Struct. Funct. Genet.* 37, 582–591.

Sorenson, J.M., and Head-Gordon, T. 2000a. Matching simulation and experiment: A new simplified model for simulating protein folding. *J. Comp. Biol.* 7, 469–481.

Sorenson, J.M., and Head-Gordon, T. 2000c. Unpublished.

Sorenson, J.M., and Head-Gordon, T. 2002. Toward minimalist models of larger proteins: A ubiquitin-like protein. *Proteins: Struct., Funct., Genet.* In press.

Villegas, V., Martinez, J.C., Aviles, F.X., and Serrano, L. 1998. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* 283, 1027–1036.

West, M.W., and Hecht, M.H. 1995. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.* 4, 2032–2039.

Address correspondence to:
Teresa Head-Gordon
Department of Bioengineering
University of California, Berkeley
472 Donner Hall
Berkeley, CA 94720-1762

E-mail: tlhead-gordon@lbl.gov